A Conceptual Model for Information Retrieval with UMLS

Michel Joubert, Ph.D., Marius Fieschi, M.D.,Ph.D. and Jean-Jacques Robert, M.S.
CERTIM. Faculté de Médecine.
Boulevard Pierre Dramard. F-13326 Marseille Cedex 15. France

Information retrieval in large information databases is a non-deterministic process which needs a sequence of search steps generally. One of the main problems to which the end-users are faced is to parse efficiently their questions into the query language that the computer systems allow. Conceptual graphs were initially designed for natural language analysis and understanding. Due to their closeness to semantic networks, their expressiveness is powerful enough to be applied to knowledge representation and use by computer systems. This work demonstrates that conceptual graphs are a suitable means to model the end-users querieson the basis of the thesaurus and the semantic network of the UMLS project.

1. INTRODUCTION

Information retrieval in large medical information databases is a non-deterministic process which needs a sequence of search steps generally. One of the main problems to which the end-users are faced is to parse efficiently their intentional questions in the query language that the computer systems allow. Several authors propose specific logics or hypertext-based technologies to assist end-users in their task [12,10]. Other authors propose the use of an expert system to assist the end-users in their searches [9]. Another powerful capability consists in providing users with customisation facilities [7,8]. The consultation of knowledge in databases is relevant from another process. Specially, a end-user may consult a knowledge base not only to capture, but also to discover information. In this last case, it is not easy, and often impossible, to define the intention of the end-user a priori. Here also the hypertext technology applies successfully [4,2]. Many authors propose methods based on semantic networks which are commonly used to express declarative knowledge relevant to concepts and their relationships, as well as procedural knowledge that represents the behavioural properties of the concepts [14].

The National Library of Medicine allows with the project UMLS a raw material composed of a thesaurus which transcends the most commonly used medical nomenclatures and thesauri, and a semantic network which describes general relationships between classes of concepts. On the basis of Meta-1 and of the semantic network of UMLS, it is possible to create a lattice of concepts. With this lattice and the relationships defined between classes of concepts a end-user is guided to create semantic networks, represented by conceptual graphs, which describe his focus on a documentary database or on a knowledge base, for example. Theoretical operations on these graphs may be applied to reduce the intrinsic ambiguity of queries expressed in natural language which must be translated in the query language of the system. This work intends to show how internal conceptual graphs built from an interactive navigation of a end-user in UMLS can be exploited to express his queries as naturally as possible.

2. THE UMLS KNOWLEDGE

The Unified Medical Language System (UMLS) is a project of the National Library of Medicine [6]. UMLS has two components which are a meta-thesaurus, called Meta-1, and a semantic network. Meta-1 contains MESH, the index thesaurus of MEDLINE, but also several medical nomenclatures among the most useful [16]. The core concepts of Meta-1 are connected to generic types of concepts in a semantic network. These types of concepts are interconnected by semantic relationships [11]. The current third experimental version of UMLS (August 1992) allows a homogeneous and more easily readable data model [17] than the other previous versions, as noticed in [18].

The data structure of Meta-1 is based on hierarchies and associations. The association relationship links a given term to related terms and to a preferred term. The (pre)order relationship structures the preferred terms into more generic terms and more specific terms. This later relation divides the thesaurus into several so-called microthesauri, according to a local specificity. For example, the term Coronary Arteriosclerosis appears twice in Meta-1: firstly as a process involved in coronary diseases viewed as heart diseases, secondly as an arteriosclerosis localised into the coronary arteries and causing arterial occlusive diseases. The presence of micro-thesauri translates the various contexts from which a same medical concept can be viewed and, thus, the complexity of the medical domain.

0195-4210/92/$5.00 © 1994 AMIA, Inc.

a. [Coronarography]->(DIAGNOSES)->[Coronary Arteriosclerosis]

b. [Coronarography]->(DIAGNOSES)->[{Coronary Arteriosclerosis, Coronary Aneurysm, Coronary Thrombosis}]

c. [Angiocardiography]->(DIAGNOSES)->[Heart Diseases]
[Coronarography]->(DIAGNOSES)->[Coronary Diseases]

d. [Coronary Artery Bypass]->(TREATS)->[Coronary Diseases]
[Coronarography]->(DIAGNOSES)->[Coronary Arteriosclerosis]
[Coronarography]->(DIAGNOSES)->[Coronary Arteriosclerosis]<-(TREATS)<-[Coronary Artery Bypass]

Figure 1: conceptual graphs notations and operations

The semantic network of UMLS associates types of medical concepts between them with semantic relationships. The types of concepts are organised in a hierarchy where, for example, Physiologic Function and Pathologic Function are children of Biologic Function, and Disease or Syndrome is a child of Pathologic Function. There are about thirty different semantic relationships. Among them, for instance, DIAGNOSES applies on the two types Diagnostic Procedure and Pathologic Function, TREATS applies on Therapeutic or Preventive Procedure and Pathologic Function. These semantic relationships are defined at a so general level that is not always possible to map a type onto the concepts linked to it with an automatic pertinent inheritance of the meaning that the relationships convey: it is obvious that every diagnostic procedure can not be used to diagnose every pathologic function. Nevertheless, since a Coronarography is linked to the type Diagnostic Procedure it can be used to diagnose a Coronary Arteriosclerosis that is linked to the type Disease or Syndrome and thus to the type Pathologic Function.

3. CONCEPTUAL GRAPHS

Conceptual graphs were initially designed for natural language analysis and understanding [13]. Due to their closeness to semantic networks and first order logic, their expressiveness is powerful enough to be applied to knowledge representation [15] and use [5]. Their many attractive features have been noted previously by medical informatics researchers [1,3]. Their properties include the ability to represent sentences as partial graphs that define conceptual structures involving individual terms, to combine partial graphs to express complex relationships between concepts, and to allow a standard means to represent knowledge which can be mapped onto other representations (e.g. database systems) or formal systems (e.g. first order logic).

Conceptual graphs are bipartite graphs involving both concepts and relationships between them. The fundamental assumption to apply the conceptual graphs theory is that concepts are organised in a lattice. Due to this fact, each of the concepts may have one or more fathers representing more generic concepts. And, each concept in the lattice may have one, and more generally, several children representing more specific (specialised) concepts. The top and the bottom of the lattice are nodes which are not involved in graph operations. On this basis, valid operations can be applied to derive new graphs from existing ones. In the following we will adopt a simplified representation of conceptual graphs elements according to our current needs. Conceptual graphs involve two kinds of nodes: conceptual nodes denoted by square brackets, and conceptual relations denoted by parentheses. For instance, the sentence "a coronarography can diagnose a coronary arteriosclerosis" is parsed by the first graph of figure 1.a. Within a node, braces denote a set of concepts, separated by commas, which express a selection constraint. For instance, the conceptual graph of figure 1.b parses that "a coronarography can be used to diagnose a coronary arteriosclerosis, aneurysm or thrombosis".

The first valid operation that applies on conceptual graphs is the specialisation which consists in the replacement of concepts in a graph by one of their specialised concepts. In other words, semantic relationships stated at a general level are inherited by more specialised levels. For example, since Coronarography is a kind of Angiocardiography, the former term is a specialisation of the later, and Coronary Diseases is more specific than Heart Diseases, the first graph of figure 1.c is specialised by the second one. A second operation that processes on conceptual graphs is the join which consists in connecting two graphs on two concepts, one of them being a specialisation of the other. For example, the two first conceptual graphs of figure 1.d can be joined, since Coronary Arteriosclerosis is a specialisation of Coronary Diseases, to produce the third graph.

4. INFORMATION RETRIEVAL WITH UMLS

The objective of this work is to assist a user to express his queries to existing medical information databases as naturally as possible. To reach this objective, it is necessary to provide the users with the tools able to customise their accesses to the information databases. A part of the customisation process consists in the capability allowed by the computer system to present to the users the terms used to index information in their semantic environment. The customisation process is completed by tools that allow users the capability to exploit this semantics interactively by a navigation from concepts to other ones in following the semantic links which connect them. The drawing of figure 2 schematises our approach. The lower level is the thesaurus used to index information in a database. The upper level is the conceptual view that a user has to the database. The intermediate level is a semantic network which structures semantically the general concepts which are instanciated by specific terms at the lower level. The lower level is constituted by Meta-1. The intermediate level is the semantic network of UMLS. Conceptual users views on the information databases are represented by conceptual graphs at the upper level. These graphs are built with the elements (types of concepts and semantic relationships) of the semantic network and the terms of Meta-1 involved by a user after he has navigated in Meta-1 through the links allowed by the semantic network. These graphs are further combined and refined by the means of graphs operations.

Since the conceptual graphs theory needs a lattice of concepts to be applied, the data model of Meta-1 must be enhanced. Meta-1 is transformed by removing the duplicated terms and linking the unique remaining term to all its direct ancestors. Each core concept is linked to one or possibly several types of concept in the semantic network. The creation of an artificial top, to which all the most generic types will be linked, and of an artificial bottom, to which all the most specific terms will be linked, transforms Meta-1 and the types of the semantic network of UMLS into a lattice. This is partially illustrated by figure 3, where the types of concepts are written in capital letters. The constraint that the concepts must be organised in a lattice being satisfied, it is now possiblr to apply the conceptual graph theory.

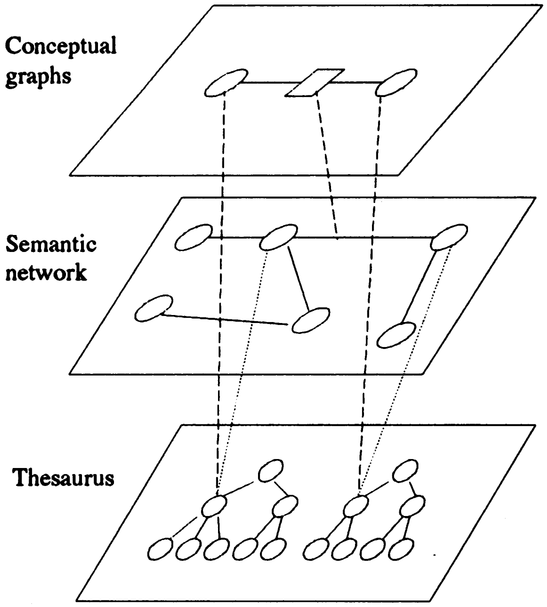

Figure 2: building conceptual graphs from a thesaurus and a semantic network

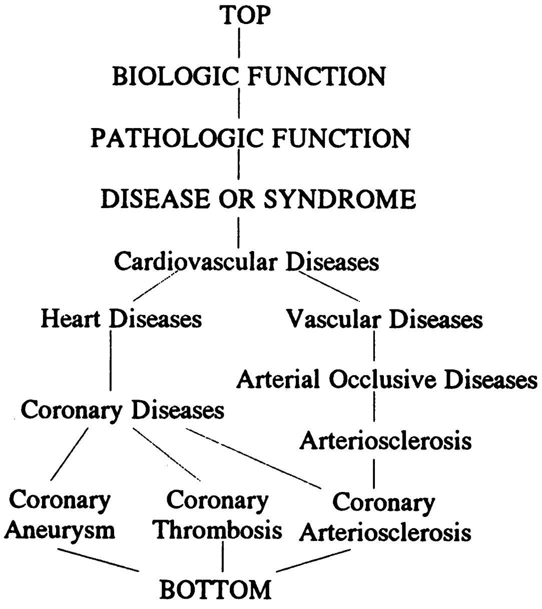

Figure 3: a part of the lattice issued from Meta-1 and the types of concepts of UMLS.

Let us consider a end-user who searches information in a medical information database. He is looking for coronary diseases, their diagnoses and treatments. After some interaction, the end-user selects, for example, the term Coronary Arteriosclerosis in the thesaurus which allows access to the database. Automatically, from the lattice of concepts, the computer system deduces

a. [Coronarography]->(DIAGNOSES)->[Coronary Arteriosclerosis]

b. [Coronary Artery Bypass]->(TREATS)->[Coronary Arteriosclerosis]

c. [Coronarography]->(DIAGNOSES)->[Coronary Arteriosclerosis]<-(TREATS)<-[Coronary Artery Bypass]

d. [Coronarography]->(DIAGNOSES)->[{Coronary Arteriosclerosis, Coronary Aneurysm, Coronary Thrombosis}]

e. [Coronarography]->(DIAGNOSES)->[Coronary Diseases]

Figure 4. conceptual graphs built by a user's navigation inside a thesaurus and a semantic network.

that this selected concept is a Disease or Syndrome which is itself a Pathologic Function. The semantic network includes several relationships applying to Pathologic Function. Thus, the computer system displays them to the user together with the other types of concepts that they connect. Let us suppose that the end-user selects the relationship DIAGNOSES and thus the target type Diagnostic Procedure which is connected to it. Since it is not possible to deduce automatically from the lot of diagnostic procedures those which are relevant from the coronary diseases, the end-user has to select by himself either a specific procedure (e.g. Coronarography) or a generic term including a set of relevant procedures (e.g. Angiocardiography). Let us suppose that the end-user has selected the term Coronarography. At this step of the process, the system internally represents the result of the navigation achieved by the user with the first conceptual graph 4.a. In a comparable way, the end-user selects the relationship TREATS that leads to the type Therapeutic or Preventive Procedure and then selects the term Coronary Artery Bypass. This is represented by the conceptual graph 4.b. The join of these two graphs produces the conceptual graph 4.c which parses the sentence "coronary arteriosclerosis diagnosed by coronarography and treated by coronary artery bypass". Thus, the system provides a user with a means to build partial graphs step by step and to join them to represent a complete sentence. This scenario shows how users views represented by conceptual graphs are built at the upper level of figure 2, after a navigation from terms to terms in the thesaurus by the means of the relationships of the semantic network.

Let us consider another scenario. A end-user has previously defined and stored in the computer memory a set of conceptual graphs that define his interest in the domain of coronary diseases. Among them are the two graphs 4.b and 4.d. Let remark that the diseases names placed between the braces in 4.d are all coronary diseases according to the lattice of figure 3. Thus, every time the user asks the system in terms of coronary diseases, the system transforms automatically the query in specialising Coronary Diseases by each of the previously stored specialisations: the conceptual graph 4.e is specialised into the graph 4.d. If, moreover, the end-user asks for treatment in the same query, the join of the later graph with the graph 4.b produces the conceptual graph 4.c. Other comparable graphs could be produced with the use of stored graphs describing the relevant treatments of thromboses and aneurysms. This scenario shows how partial graphs are stored after they have been produced as in the first scenario and are used to describe the user's interest in the information database. These graphs will serve to precise the queries that the user in now able to express more easily than he should have to explicit the query completely. The consistency of the query produced by the means of these graphs is guaranteed by the fact that they are issued from the knowledge of the system.

5. DISCUSSION AND CONCLUSION

The conceptual model presented in sections 3 and 4 is a suitable means to represent semantically the result of an interaction with a end-user after he has navigated from terms to terms in the thesaurus by the means of the relationships of the semantic network. The conceptual graph resulting from this interaction is further parsed into the query language of the information server according to the precision of the concepts that it allows. Conceptual graphs are used to represent users views also and are thus a suitable tool for query control and assistance since the valid operations on both end-users dynamically built graphs and previously stored graphs guarantee the coherence of the results.

Due to the presence of various points of view on medical concepts, the structure of a thesaurus is often complex. Thus, the users need to customise their accesses to information databases according to their own views on medical concepts. Section 4 has demonstrated how the combination of partial conceptual graphs issued from users interactions with a thesaurus augmented by a semantic network parses the interest of users in a medical information database. However, we must be careful to not introduce inconsistency between the customised knowledge and the system knowledge. Users partial conceptual graphs are, for the moment, built with concepts and relationships issued from the semantic network. In a next future, we will introduce users concepts described by the means of definition relationships linking them to concepts already present in the system knowledge. This capability will have to be accompanied by a consistency control between the used definition relationships and the semantic relationships that link the concepts in the semantic network.

The project UMLS of the National Library of Medicine provides a suitable raw material which is a powerful start point for investigation in the domain of users interfaces with medical information databases. It permitted us to experiment successfully the use of conceptual graphs for searching in large information databases.

ACKNOWLEDGEMENTS

This work has been partially founded by the *Direction de la Recherche et des Etudes Doctorales* (DRED) of the French Ministry of Education and by the *Délégation à l'Information Scientifique et Technique* (DIST) of the French Ministry of Research and Technology. The authors thank the National Library of Medicine which graciously furnished us the UMLS material.

REFERENCES

[1] J. Bernauer. Conceptual Graphs as an Operational Model for Descriptive Findings. Proc. 15th SCAMC. P.D. Clayton editor. McGraw-Hill, 1991: 214-218.

[2] P.D. Bruza, T.P. van der Weide. Stratified Hypermedia Structures for Information Disclosure. The Computer Journal 35, 1992: 208-220.

[3] K.E. Campbell, M.A. Musen. Representation of Clinical Data Using SNOMED III and Conceptual Graphs. Proc. 16th SCAMC. M.E. Frisse editor. McGraw-Hill, 1992: 354-358.

[4] M.E. Frisse. Searching for Information in a Hypertext medical Handbook. Comm. ACM 31, 1988: 880-886.

[5] R.T. Hartley, M.J. Coombs. Reasoning with Graph Operations. In: Principles of Semantic Networks: explorations in the representation of knowledge. J.F. Sowa editor. Morgan Kaufmann, 1992: 487-505.

[6] B.L. Humphreys, D.A.B. Lindberg. Building the Unified Medical Language System. Proc. 13rd SCAMC. L.C. Kingsland editor. IEEE Computer Society Press, 1989: 475-480.

[7] M. Joubert, D. Riouall, M. Fieschi, G. Botti, H. Proudhon. Contextual Aids for Medical Information Retrieval. Proc. MEDINFO 92. K.C. Lun, P. Degoulet, T.E. Piemme, O. Rienhoff editors. North-Holland, 1992: 1522-1527.

[8] M. Joubert, M. Fieschi, J-J. Robert, G. Botti. Customization of Medical Information Retrieval. Proc. MIE 93. To appear.

[9] L.C. Kingsland, E.J. Syed, D.A.B. Lindberg. Coach: an Expert Searcher Program to Assist Grateful Med Users Searching MEDLINE. Proc. MEDINFO 92. K.C. Lun, P. Degoulet, T.E. Piemme, O. Rienhoff editors. North-Holland, 1992: 382-386.

[10] D. Lucarella. A Model for Hypertext-Based Information Retrieval. Proc. European Conf. on Hypertext. A. Rizk, N. Streitz, J. André editors. Cambridge University Press, 1990: 81-94.

[11] A.T. McCray. The UMLS Semantic Network. Proc. 13rd SCAMC. L.C. Kingsland editor. IEEE Computer Society Press, 1989: 503-507.

[12] C.J. van Rijssbergen. Towards an Information Logcic. Proc. 12th ACM SIGIR Conf. on Research and Development in Information Retrieval. N.J. Belkin, C.J. van Rijsbergen editors. 1989: 77-86.

[13] J.F. Sowa. Conceptual Structures - information processing in mind and machine. Addison Wesley, 1984.

[14] Principles of Semantic Networks: explorations in the representation of knowledge. J.F. Sowa editor. Morgan Kaufmann, 1992.

[15] J.F. Sowa. Conceptual Analysis as a Basis for Knowledge Representation. Tutorial hand-book. MEDINFO 92.

[16] M.S. Tuttle, D.D. Sheretz, M.S. Erlbaum, N. Olson ; S. Nelson. Implementing Meta-1: the First Version of the UMLS Metathesaurus. Proc. 13rd SCAMC. L.C. Kingsland editor. IEEE Computer Society Press, 1989: 483-487.

[17] M.S. Tuttle and all. The Homogenization of the Metathesaurus Schema and Distribution Format. Proc. 16th SCAMC. M.E. Frisse editor. McGraw-Hill, 1992: 299-303.

[18] Y. Yang, C.G. Chute. A Schematic Analysis of the Unified Medical Language System. Proc. 15th SCAMC. P.D. Clayton editor. McGraw-Hill, 1991: 204-208.